Effect of user sessions on the heuristic usability method

International Journal of Open Source Software and Processes
Volume 9 • Issue 1 • January-March 2018

Effect of User Sessions on the Heuristic Usability Method

Jehad Alqurni, School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, UK

Roobaea Alroobaea, College of Computers and Information Technology, Taif University, Taif, Saudi Arabia

Mohammed Alqahtani, College of Computer Science and Information Technology, Imam Abdulrahman Alfaisal University, Dammam, Saudi Arabia

ABSTRACT

Heuristic evaluation (HE) is a widely used method for assessing software systems. Several studies have sought to improve the effectiveness of HE by developing its heuristics and procedures. However, few studies have involved the end-user, and to the best of the authors' knowledge, no HE studies involving end-users with non-expert evaluators have been reported. Therefore, the aim of this study is to investigate the impact of end-users on the results obtained by a non-expert evaluator within the HE process, and through that, to explore the number of usability problems and their severity. This article proposes introducing two sessions within the HE process: a user exploration session (UES-HE) and a user review session (URS-HE). The outcomes are compared with two solid benchmarks in the usability-engineering field: the traditional HE and the usability testing (UT) methods. The findings show that the end-user has a significant impact on non-expert evaluator results in both sessions. In the UES-HE method, the results outperformed all usability evaluation methods (UEMs) regarding the usability problems identified, and it tended to identify more major, minor, and cosmetic problems than other methods.

KEYWORDS

Heuristic Evaluation, Inspection Usability, Non-Expert Evaluator, Usability, Usability Testing, User Session

1. INTRODUCTION

A revelation in technologies has led to a significant spread in system products and has thus increased the demand for system product development. One of the most popular types of system products is web-based systems (Sova and Nielsen, 2003), which play a significant role in enabling private or public organizations to provide information and services to end-users (Harrison and Petrie, 2007) (Alqurni and Pooley, 2016). An end-user is anyone who can use the target system and interact with its interface (ISO 9241-11, 1998), (Alqurni and Pooley, 2016). The user interface of web-based systems is the mediator of the interaction between the end-user and the website. The usability of the user interface has thus increasingly attracted interest in our world because of the growing increase in the number of users every year. Quantitatively, the number of users of web-based systems has dramatically grown from 1,971 million users (28.8% of the world population) in 2010 to 3,675 million (50.1% of the world population) in 2016 (Group, 2016).

DOI: 10.4018/IJOSSP.2018010104

Chen (2012) stated that the success or failure of a product is significantly affected by its usability. Nielsen (2001) also reported that the reason behind the abandonment of 50% of sales websites is poor website usability. Commercial sites have also been shown to experience difficulties in the competitive environment due to poor usability (Osterbauer, Kφhle, Grechenig, & Tscheligi, 2000). For example, nearly 39% of online buyers were found to have failed to accomplish their purchases online due to usage difficulties (S. Y. Chen & Macredie, 2005). This shows that good or poor usability plays a pivotal role in the success or failure of website products. Consequently, usability is considered one of the most important factors that influence the user interface of a web-based system, and it plays a significant role in fulfilling users' satisfaction. Therefore, several usability evaluation methods (UEMs) have been developed to measure the level of usability. The most common UEMs used in web-based systems are the usability testing (UT) and heuristic evaluation (HE) methods (Fernandez, Insfran, & Abrahão, 2011). Although the HE method is described as being more affordable than the UT method, its results are prone to the opinion of the evaluators. UT results are derived from the end-user and are thus the consequence of real problems and it is limited to user tasks.

Several studies have attempted to improve the effectiveness of the HE method by examining the major factors of HE such as lists of heuristics or evaluator expertise. Expertise is one of the most important factors contributing to the improvement of the HE (Hwang & Salvendy, 2007) method, which can be used by either expert or non-expert evaluators. Nielsen (1992) has described the non-expert evaluator as one who lacks experience both in usability and in the system domain but who have a solid background in computer field. In contrast, expert evaluators are described as those who have expertise in usability. Although the former yield more accurate results (Nielsen, 1992), several studies indicated that expert evaluators are difficult to find (Äijö & Mantere, 2001; Desurvire & Thomas, 1993; Nielsen, 1999; Paz, Paz, Villanueva, & Pow-Sang, 2015). In addition, Fernandez et al. (2011) stated that: "Although inspection methods are intended to be performed by expert evaluators, most of them were applied by novice evaluators such as Web designers or students."

Regarding the role of the end-user in the process of the HE method, most studies on the HE method depend on Nielsen's (1993) argument that HE is not performed by end-users but by expert or non-expert evaluators. Fernandez et al. (2011) stated that one of the main disadvantages of HE is that it does not involve the end-user. In contrast to Nielsen's view, Muller, Matheson, Page, and Gallup (1998) developed an HE method by involving users as "web-domain experts" as part of the evaluation team, and the technique then becomes that of Participatory Heuristic Evaluation (PHE), by combining experts with users. The authors claimed that if users are easy to recruit, then PHE can be as cost-effective as traditional HE. However, users in this method are described as "web-domain experts," which differ from real end-users. In addition, this approach still requires expert evaluators on the PHE team.

Thus, continued research is necessary to study and address the usability topic, with emphasis on the exploration of the key factors of the HE method. Few studies have investigated the role of end-users in a HE method such as that used by Muller et al. (1998). In fact, the absence of the user's point of view has been described by many researchers as one of the shortcomings of the HE method, (Holzinger, 2005; Oracle, 2012; Zaharias & Koutsabasis, 2012). Despite the effectiveness of the HE method of using evaluators as simulated users, Fu, Salvendy, and Turley (2002) stated that the evaluator does not represent the real user of the system. Therefore, the evaluator may fail to simulate the real user of the system in two ways: by failing to detect potential problems, or by identify usability problem which are ultimately not considered a real problem for the user. (Matera et al., 2006). It is clear that there is insufficient research into the involvement of users with the evaluators within HE method that enables exploration of both points of view. Hollingsed and Novick (2007) support this conclusion by stating, "It remains an open issue as to why usability professionals, in practice, rely on single-perspective methods, typically involving users, or experts, but not both." To date, to the best of our knowledge, there have been no studies with a focus on the influence of end-users on non-expert evaluator output within the HE method. Thus, the main aim of this research is to investigate

the impact of end-users on non-expert evaluator output regarding the number of usability problems, their severity, and usability performance metrics (thoroughness, validity, and effectiveness), in cases in which end-users are easy to recruit, and expert evaluators are difficult to find.

Based on the above, the overall aim of this research is to investigate the impact of the end-user on non-expert judgment within the heuristic evaluation method through different types of user sessions. This research poses the following question based on the primary aim of this study:

Do the proposed methods "UES-HE and URS-HE", the traditional HE method, and UT method differ in terms of the usability performance metrics (effectiveness, efficient, thoroughness), the number of usability problems detected and the severity of usability problems?

This paper is organized as follows: Section 2 provides an overview of the methods used in this research. Section 3 describes the research methods utilized in this research. Section 4 presents the results and discussion, including a comparison between the proposed methods and two solid benchmarks (HE and UT); and the last section concludes this paper.

2. USABILITY EVALUATION METHOD

The broadly used definition of usability is based on a definition by the International Organization for Standardization (ISO) as "…the extent to which the product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use ISO 9241-11." (1998). The purpose of using the usability evaluation is to ensure the quality requirements of any software product. Usability evaluation methods are a set of techniques that are used to measure the usability level of any software product. These techniques can be classified into three categories depending on the role of the evaluators: testing, inspection, inquiry, and metrics-based (Zhijun, 2007). Among the usability evaluation methods, the HE and UT methods are considered to be the most widely employed usability methods, according to a survey (Sherman, 2009). The main difference between them is that HE depends on evaluator expertise to predict potential usability problems, whereas UT depends on observing end-users to find usability problems. However, there is still no agreement as to which is the best technique, as none of the two techniques can identify all the usability problems and each of them has its advantages and disadvantages. Regarding UT methods, the primary advantage involves recruiting potential end-users and the primary disadvantage is that these methods are expensive and time-consuming. As for the HE method, the main benefits are that it is less expensive, and it requires fewer resources than other methods and is, therefore, less time-consuming to implement than usability testing. However, the primary disadvantage is that it does not involve end-users and it is a subjective assessment (different evaluators can produce different results), and it depends on the evaluators' experience. The use of both methods can be expected to offer improved results but would be costly. Therefore, it is important to improve the effectiveness of one of these methods, and in practice this will be HE method. As Baker, Greenberg, and Gutwin (2001) stated, *"Heuristic evaluation (HE) is a widely accepted discount evaluation method for diagnosing potential usability problems in user interface and HE is popular with both researchers and industry."*

2.1. Usability Testing

The UT method is considered one of the most common usability evaluation methods for ensuring quality in terms of website effectiveness. The UT method involves using an observer to observe end-users while they are performing their tasks with the aim of extracting usability problems directly from the end-user. Nielsen (1993) clarified the role of the observer as "…interpreting the user's actions in order to infer how these actions are related to the usability issues in the design of the interface." Furthermore, in contrast, he noted that HE does not require interpretation of the evaluator's actions but only collects their comments about the user interface.

Among the different UT techniques, the think-aloud protocol is described as the most valuable technique among usability evaluation methods (Holzinger, 2005). This technique simply means that the user is verbalizing their thoughts while they are using the system (Nielsen, 1993). The think-aloud protocol consists of three types of interaction techniques: concurrent, retrospective, and constructive interaction (Dumas & Redish, 1999; Nielsen, 1993; Van den Haak, de Jong, & Schellens, 2004). Concurrent interaction involves the user verbalizing their thoughts when performing a list of tasks. The results obtained by M.J. Van den Haak et al. (2004) showed that concurrent interaction is more easy to distinguish than the other two methods. The retrospective technique proceeds through two steps where the user first silently interacts with the system, after which the user talks about their opinion of the interaction. Constructive interaction is also known as "co-discovery learning" (Kennedy, 1989). This method involves two users working together to perform their tasks while verbalizing their thoughts in practice. However, the retrospective technique is described as being less frequently used whereas the concurrent technique has been pointed out as being the most common (Van den Haak et al., 2004). When the user verbalizes their thoughts, it helps the observer to understand user behavior in terms of their view about the system and the difficulties they encounter with the system (Holzinger, 2005; Nielsen, 1993). However, end-users consider the think-aloud technique to be unnatural and that it does not allow them to act naturally as they would act in a real-life situation (Nielsen, 1993; Rubin & Chisnell, 2008; Maaike J van den Haak & de Jong, 2005).

A user task is a core factor of the UT method. The list of user tasks should cover the main functions of the website. Dumas and Redish (1999) recommended that formulated tasks should be short and clear and in the users' language. Although the UT method has been implemented by requiring different numbers of end-users to perform the list of tasks, there is no agreement about the optimal number of users. Nielsen (2000) suggested that five users are sufficient to find 85% of usability problems. In addition, he stated that more than five users waste resources and time. In line with Nielsen, Virzi (1992) confirmed that five users were able to reveal 80% of usability problems. Likewise, the findings of Turner, Lewis, and Nielsen (2006) showed that three to five users are sufficient to discover most of the usability problems. In contrast, (Faulkner, 2003) argued that five users may find 55% of all usability problems. Lindgaard and Chattratichart (2007) supported this argument and claimed that five users were only able to find 35% of the total number of usability problems. In regard to using the UT method for benchmarking, (Nielsen, 2006) recommended using 20 users.

2.2. Heuristic Evaluation

Heuristic evaluation is a usability inspection method, and it requires a small set of evaluators to use a set of usability principles (known as "heuristics"), to examine the user interface to predict potential usability problems. These heuristics are the visibility of the system status, the extent to which the system corresponds to the real world, user control and freedom, consistency and standards, error prevention, recognition rather than recall, flexibility and efficiency of use, aesthetic and minimalist design, the extent to which the system helps users recognize, diagnose, and recover from errors, and the help and documentation offered by the system (Nielsen, 1994a). This list of heuristics is used as an aid tool for evaluators during inspection of the website to remind them about the area of usability issues.

Although there is no agreement about the way to use the HE method, Nielsen (1994b) recommendations are to consider it as a typical method of evaluation procedures. These procedures have been grouped into three main sessions. First, a pre-evaluation/training session can be conducted before starting the actual evaluation. This session is very useful for non-expert evaluators more than for expert evaluators. Its usefulness is that it can help evaluators who lack experience with the domain of the target website to enable these evaluators to familiarize themselves with the website domain. In addition, it contributes to familiarize evaluators with the list of heuristics used and provides training with respect to the steps and procedure of evaluation. Second, the actual evaluation session is the main session, where each evaluator performs the evaluation independently without communicating with the other evaluators to ensure there is no bias between them. This session can be conducted

in two phases. The first phase is to navigate through the target website to obtain a general feeling about the entire website, whereas the second phase is more focused on inspecting usability problems in each particular part of the website. During the actual evaluation session, evaluators are required to list usability problems as best they can. Next, a debriefing session follows the actual evaluation session and allows evaluators to meet to discuss their results. The technique used in this session is brainstorming. Finally, Nielsen also recommended the combination of all the evaluators' reports of usability problems into one list and the estimation of the severity of each problem.

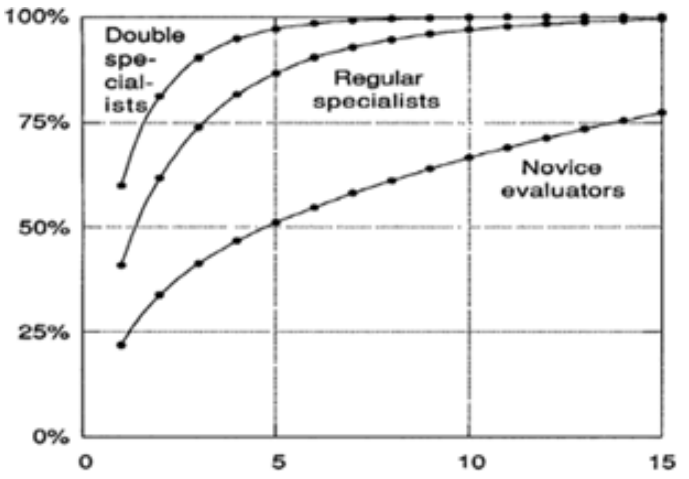
Evaluator expertise is the most important factor among various factors affecting the HE method and its results. In an early study, Nielsen (1992) classified evaluators according to their expertise as follows:

- Novice/non-expert evaluators, who lack experience both in usability and in the system domain, but who have a solid background in the field of computers;
- Regular usability specialists, who have expertise in usability but not in the domain of the interface;
- Double usability specialists, who have experience in both usability and the domain of the interface.

There is also no consensus regarding the adequate number of evaluators to use the HE method. Regarding the number of evaluators, it depends mainly on the type of evaluator experience. In an early study, Nielsen (1992) estimated that finding 75% of the system problems would require two to three double usability specialists. In contrast, three to five regular usability specialists would be able to find the same percentage of problems, whereas fourteen novice/non-expert evaluators would be required. This study indicated that five and eight novice/non-expert evaluators would be sufficient to find 51% and 60% of the problems, respectively, as shown in Figure 1.

Although expert evaluators achieve more satisfactory outcomes, these evaluators are difficult to find and may not be available. Therefore, some studies have investigated the effectiveness of non-experts (Äijö & Mantere, 2001; Botella, Alarcon, & Peñalver, 2013; Fernandes, Conte, & Bonif'cio, 2012; Howarth, Smith-Jackson, & Hartson, 2009; Koutsabasis, Spyrou, Darzentas, & Darzentas, 2007; Slavkovic & Cross, 1999). Despite the benefits of the HE method, obtaining results relies on the skill and expertise of evaluators to assess the website interface. This leads to the opinion of the end-user being ignored during this evaluation process. Thus, complete dependence on this method means that the evaluators are trying to play the end user's role by trying to perceive the problems they

Figure 1. Proportion of usability problems found by evaluators depending on their expertise (Nielsen, 1992)



may encounter while using the end-user interface instead of involving them directly in the evaluation process. To date, there has been little current research to examine the end user's role in the evaluation process, and this is the main aim of our research.

3. RESEARCH METHODS

The above-mentioned research question was tested by using an experimental approach. This experiment investigated the causal relationship between independent and dependent variables. Before the data are analyzed and the results discussed, some steps of the research procedures are described as follows.

3.1. Research Design

Between-subject experimental design was adopted to examine two or more UEMs where a different participant was recruited for each UEM of the independent variable. These variables are UEMs (UES-HE, URS-HE, HE, and UT methods). The dependent variables are the number of usability problems, severity rating, and the usability performance metrics (thoroughness, validity, and effectiveness).

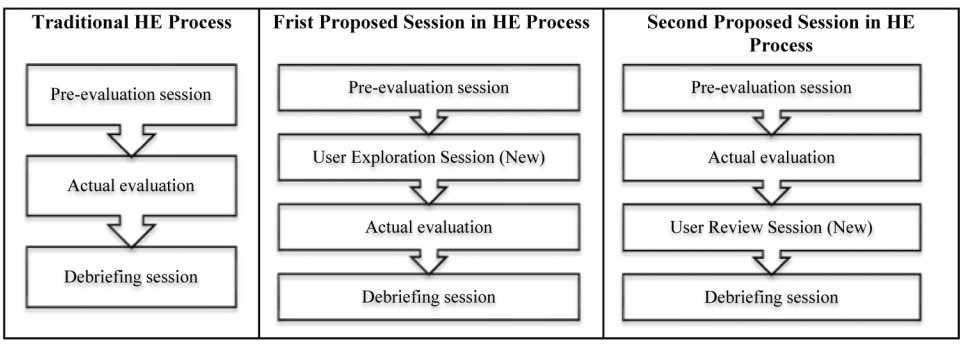
3.2. Evaluation Process

As there is no particular procedure for performing the HE process, this research adopted Nielsen's (1994b) recommended approach for implementing the HE method. These processes can be summarized as several stages including three main sessions. The first stage is a pre-evaluation session, to familiarize the evaluator with the HE processes. Then, in the actual evaluation session, each evaluator examines the interface independently and identifies the expected problems, according to a short list of heuristic guidelines. However, evaluators are not limited to a short list of heuristic guidelines; instead, they can record any usability issue that is not relevant to any heuristics list. Finally, the single results of each evaluator are combined to a single list throughout a debriefing session. The UES-HE and URS-HE processes are the same as the HE proses except that a new proposed user session was included in both methods as described in the next section.

3.3. Proposed Sessions

The main aim of this study is to investigate the impact of end-users on non-expert evaluator results through the HE method. The impact on the results is clearly seen in the number of usability problems discovered. Our approach to achieve this aim was to propose two user sessions to examine the influence of end-users on the results obtained by non-expert evaluators, as illustrated in Figure 2:

Figure 2. Proposed sessions



- **User Exploratory Session of Heuristic Evaluation method (UES-HE):** The non-expert evaluators used a "question-asking protocol" with the end-users to understand whether they experienced difficulties with understanding and using the system. In this session, the users performed tasks, and the evaluator tried to encourage them to complete the tasks with direct questions while noting any issues. In addition, the evaluator asked additional in-depth questions to understand any issues the task did not cover;
- **User Review Session of Heuristic Evaluation method (URS-HE):** Evaluators aimed to minimize potential problems the user had not yet experienced at that time or tried to modify existing issues by reviewing the potential usability problems list with the end-user by using a "free discussion protocol." The evaluator was not restricted to gathering the users' impressions about every usability issue, but could expand the review to discuss other usability issues that may have emerged during the review session to add new problems.

Each usability problem was classified depending on its severity rating. Nielsen (1994b) recommended using the following severity ratings to estimate the severity of each problem:

1. Cosmetic problem only – need not be fixed unless additional time is available;
2. Minor usability problem – fixing this should be given low priority;
3. Major usability problem – important to fix, so should be given high priority;
4. Usability catastrophe – imperative to fix this before the product can be released.

With regard to the severity of the problem, and to obtain more reliable results, it is recommended the severity of the problem be determined by an independent evaluator. Nielsen and Mack (1994) recommend using "…the mean of the severity judgments from several evaluators to obtain much more reliable results." In this study, two independent experts were recruited to score the severity of usability problems according to the four categories described above. Consequently, independent evaluators, who are the second and third authors of this article, have classified each problem based on their experience in the usability field.

Several lists of principles or guidelines have been published for user interface evaluation (AlRoobaea, Al-Badi, & Mayhew, 2013; Shneiderman, 1998). Despite these proposed lists of heuristics, Nielsen's usability heuristics are the most widely used and their usefulness has already been studied (Allen, Currie, Bakken, Patel, & Cimino, 2006; Edwards, Moloney, Jacko, & Sainfort, 2008; Lin Chou & Mustafa, 2014; Paz et al., 2015). In addition, Lin Chou and Mustafa (2014) described Nielsen's 10 usability heuristics as being "more concise and easily understood." However, the evaluator is not restricted to finding only those usability problems related to these heuristics but can also include any usability issue noted during the evaluation and not related to these heuristics.

3.4. Target Website and Participants

Web-based systems are currently the most widely used by people around the world. Both governments and organizations use these systems as an effective approach to provide information and necessary services for people. The rapid growth of web-based systems has resulted in several types of websites being developed.

Dynamic websites are considered as one of the most rapidly spreading types of web-based systems. To achieve our research aims, we concentrated on dynamic websites and excluded other kinds such as commercial and static websites for several reasons: the end-user should be able to use the website free of charge; the website should consist of various types of functionalities, target a broad range of end-users, and provide public information and services. The target website used in this research is designed to provide general users with information about tourism events and activities. This website is "Saudi Events" website (http://www.saudievents.sa), which is provided by the Government of Saudi Arabia. It has been selected based on the selection criteria mentioned above.

All the non-expert evaluators recruited into this study were computer science students who have completed two years at undergraduate level. End-users were recruited from the general community because the targeted website targets the community in general and does not target a segment with distinctive characteristics. Characteristics, such as age and experience, were taken into account in distributing the participants to each method-group. Also, it should be noted that none of the participants had ever used the target website.

This research recruited three groups representing non-expert evaluators for the UES-HE, URS-HE, and HE methods. Also, it recruited one group representing potential end-users for the UT method. First two groups representing the two proposed user sessions (UES-HE and URS-HE) comprised 16 non-expert evaluators who were recruited for each method. To achieve the research aim of knowing the extent to which the user influences the non-expert evaluator within the traditional HE method, eight end-users were involved in each of these two approaches (UES-HE and URS-HE) to represent the user session in each method. Because UES-HE was included in the first proposed method, each two evaluators observed one user while user interacting with the targeted system before the evaluators conduct the actual evaluation session. So, the number of users was 8 versus 16 evaluators. This is also the case in the second proposed method, but the difference is that the evaluators joins the users after the actual evaluation session and that is why this method is called the review session.

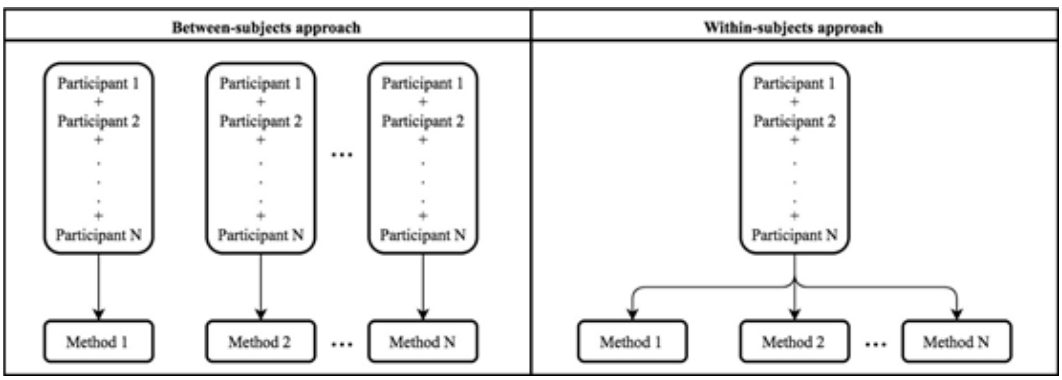
Sixteen evaluators were recruited for the third group, representing the HE method as the first solid benchmark. The last group represented the UT method and consisted of 20 users. The research results were validated using both the HE and UT methods as benchmarks because they are the most solid benchmarking methods in the field of usability research.

3.5. Experimental Approach

This research as other experimental research that must define the approach used to assign participants to the experimental conditions. The two common ways to establish the research environment are between-subject and within-subject designs. To clarify, between-subject designs are used to examine the difference between individuals or groups of participants in research. Notably, a comparison of the groups reveals the effect of the treatment given. On the other hand, within-subject designs, the participants are exposed to more than one treatment at a time. The distribution of participants for each design type is shown in the Figure 3.

More so, there are several factors to be considered when determining the aspects to assign participants in UEM. The main advantages of the between-subjects approach are that it is a clear approach, can be implemented in less time and it reduces the frustration of participants while they are performing UEMs. However, it needs a more significant number of participants. In contrast, the within-subjects approach is characterized by preventing individual differences between groups of

Figure 3. Between-subjects and within-subjects approach



participants as well as requiring a few number of participants. Nevertheless, this approach has been criticized because it is exposed to learning effect and it consumes much time from participants which may cause frustration among participants.

Considering the context of the questions to be answered, the between-subjects design was chosen as the most appropriate approach. It is worth noting that the experiment requires the participants to perform more than one session during the implementation of the UEM. Consequently, measures should be put in place to ensure the participants are not frustrated by the process as this may result in the withdrawal of some of them from experiment or may affect their behavior due to fatigue.

3.6. Data Analysis and Performance Metrics

The quality of the results produced by each method was determined by conducting a comparative usability evaluation between the results of the UEMs in term of usability numbers and their severity. In addition to the systematic comparison between the results of the UEMs, thoroughness, validity, and effectiveness were used as performance metrics. Thoroughness was defined as "…the extent to which a usability evaluation method can identify real usability problems…" (Khajouei, Hasman, & Jaspers, 2011). It can be calculated by finding the ratio between the number of real problems that are found to the number of problems that actually exist (Hartson, Andre, & Williges, 2003; Sears, 1997). The validity is defined as "…the extent to which a usability evaluation method accurately identifies usability problems…" (Khajouei et al., 2011). It is calculated by finding the ratio between the number of real problems that are found to the total number of issues identified as being problematic (Hartson et al., 2003; Sears, 1997). The effectiveness is defined as "…the product of thoroughness and validity…" and can be calculated by multiplying the thoroughness with the validity (Hartson et al., 2003). Thus, these metrics can be shown as follows:

Validity= (No. of real problems found by target method) / (No. of issues found by target method)
Thoroughness= (No. of real problems found by target practice) / (No. of real problems found by all methods)
Effectiveness= Thoroughness × Validity

In addition, Woolrych, Cockton, and Hindmarch (2005) stated that "Standard measures of inspection method effectiveness are calculated from the hits, misses and false positives (thoroughness = hits / (hits + misses), validity = hits / (hits + false positives), effectiveness = thoroughness × validity." All the issues identified by observers using UT methods are real problems because the process is driven by real end-users. In contrast, all usability issues detected by evaluators during the process of UES-HE, URS-HE, or HE methods are predicted problems because they are produced based on the evaluator's experience and represent the subjective judgments of the evaluator. Therefore, some of these predicted problems may not hinder the end-users during real usage and are coded as "false positives." These kinds of problems should eventually be removed from the evaluator report. In this case, the realness of a usability problem becomes more important with methods that depend on subjective judgments of the evaluator. Expert review and judgment are one of the techniques that can be used to determine the realness of a usability problem; in other words, to determine whether it is a real or false problem (Hartson et al., 2003). Therefore, two independent evaluators were involved in taking this role depending on their experiences. On the other hand, usability problems that are found when using the HE and UT methods are considered real problems and coded as "hits." Finally, real problems that are found in one method but are not identified by another method are coded as "unique problems." Ultimately, each inspection usability method aims to have a high hit rate and low false positive rate.

3.7. Validity

Validity is essential to understand that is one of the leading elements of any research. According to Preece, Sharp, and Rogers (2015), the term validity is "…concerned with whether the evaluation

method measures what it is intended to measure. This encompasses both the method itself and the way it is implemented." Apparently, it is noted that there is no specific consensus among scientists regarding the types of validity. In addition, it is noted that there are no specific measures for determining the validity of a study. However, Gray and Salzman (1998) pointed out the most common types that are relevant to the study of HCI and offered advice on how to deal with such experimental threats. These are: internal validity, construct validity, and statistical validity.

3.7.1. Internal Validity

Apparently, researchers have failed to find the specific approaches to successfully measure this type of validity. However, Gray and Salzman (1998) explained that this type of validity requires proper planning for three research elements which are: setting, selection and instrumentation. Regarding setting, they pointed out that the sample of participants should be subject to the same conditions and in the same place, regardless of the different groups in which they were appointed. Also, they stated that participants should be randomly assigned to each test group to address validity threats related to the selection of participants. Finally, they recommended that research instrumentation should not be biased toward a UEM at the expense of other UEMs. As a matter of fact, this is achieved by employing independent experts who are not participants in the experiments to assess the severity of usability problems. Also, the process of identifying and classifying usability problems must be uniform.

3.7.2. Construct Validity

Construct validity is concerned with ensuring that the research measures what it claims to measure. Therefore, all steps and procedures followed during the research should be detailed. This factor enables practitioners to understand and apply the same procedures to those proposed UEMs from the researcher. More so, Gray and Salzman (1998) also recommend avoiding the involvement of participants for more than one method of research. Failure to follow this recommendation will affect the behaviour of the participants. To clarify, this is to be observed because their performance of the first method, for example, will affect their performance in the following method. Therefore, they recommend using the between-subject whereas each group of UEMs is performed by a different group of participants.

3.7.3. Statistical Validity

Statistical validity is used to ensure that there is a significant difference between the results of each group of participants for each UEM method. Potential threats to statistical validity are represented in two aspects which are low statistical power and random heterogeneity of participants. According to Gray and Salzman (1998), "…low statistical power may cause true differences not to be noticed; random heterogeneity of participants may cause noticed differences not to be true." They offered a simple solution to address such potential threats by ensuring that the numbers of participants in each group of UEM are sufficient to conduct statistical tests. Regarding usability studies, Sova and Nielsen (2003) recommended that between 10 and 12 participants per UEM group are sufficient to conduct statistical tests that test the statistical differences between these groups.

In essence, all above recommendations were taken into consideration in advance during the preparation of the research design to ensure the validity of this research. Taking this into consideration, the statistical analysis which shows the inferential statistics of this study will be explained in detail as part of the future research.

4. RESULTS AND DISCUSSIONS

Comparative results that were obtained when investigating the quality of research results were conducted in three phases. The first phase involves presenting the result of the UT method, which can help to identify real problems in the target system. The second phase is to compare the results of

the proposed methods with the results obtained with the traditional HE method. Finally, the results of all UEMs are compared to find the best performance among them. The overall aim of any UEM is to find as many usability problems as possible. Thus, this research explores the number of usability problems for each method and their severity in the first and second phases. In respect of providing a comprehensive comparison, the last phase discusses the number of usability problems, severity rating, and usability performance metrics for all UEMs. To decide whether one method can discover adequate numbers of real usability problems, overlapping and unique problems are presented in the last phase. These phases are presented below.

4.1. Phase 1: Results of UT Method

The analysis of this research starts with UT practice. All usability problems identified by the UT method are considered as real problems because they are derived directly from the end-user. The main benefits of this list confirm the real problems that are discovered by using other methods (UES-HE, URS-HE and HE). This section presents the number of usability problems and their severity ratings. This method discovered 43 usability problems in the list of problems and all of them are coded as "real problems." In regard to the severity rating for usability problems, Table 1 provides the number of cosmetic, minor, major, and catastrophic usability problems. The UT method tended to find a larger number of major problems 37.2% (16 out of a total of 43 problems) compared to the other types of problems on the list. The method found approximately one quarter of the problems on the list, i.e., the percentage of cosmetic and minor problems was 23.3% and 25.6%, respectively, whereas the proportion of catastrophic problems was 14.0%.

4.2. Phase 2: UES-HE, URS-HE, and HE Results

The second phase entailed comparing the results of the proposed methods (UES-HE and URS-HE) with the results of the traditional HE method. Any problem identified with this approach that matches any problems found with the UT method are considered as real problems, whereas the remaining list of problems was decided by independent evaluators to determine whether they are real or false problems depending on their experience (Hartson et al., 2003). This section starts by identifying the number of usability problems and the severity rating for each method. Then, the results of the overlapping and unique problems, which can help to explore whether one method alone would be able to find an adequate number of problems, are presented.

The total number of unique problems discovered by UES-HE, URS-HE, and HE, is 64 real problems. Table 2 indicates that the UES-HE method offers satisfactory results. This method discovered 50 real usability problems, which is equivalent to 78% of the entire usability problems discovered by all three methods (64 real usability problems); thus, it outperforms the other two methods. Both URS-HE and HE methods offer unsatisfactory results. The URS-HE method discovered 35 problems, and this is equal to 56% of the total problems revealed. On the other hand, the HE method revealed only 23 usability problems, which is equal to 37.5% of the total usability problems.

Table 1. Usability problems and severity rating of UT method

	UT Method	
1: Cosmetic	10	23.3%
2: Minor	11	25.6%
3: Major	16	37.2%
4: Catastrophic	6	14.0%
Total real problem	43	100.0%

Table 2. Usability methods and discovered problems

	UES-HE Method	URS-HE Method	HE Method
No. of revealed problems	50	36	24
% of revealed problems	78%	56%	37.5%

Regarding the severity rating for the discovered usability problem by the three methods. Table 3 shows that the UES-HE method was able to discover more minor and cosmetic problems compared to the other two methods, and it was also able to discover the similar number of major problems as those found by the URS-HE and HE methods together. In terms of catastrophic problems, the result was close, where the UES-HE, URS-HE, and HE methods identified 5, 4, and 3 catastrophic problems, respectively.

4.3. Phase 3: Comparing all UEMs (UES-HE, URS-HE, HE, and UT Results)

After presenting the results of all UEMs in terms of the number of usability problems and severity rating as mentioned above, this phase of presenting the research results involves comparing the results of the three methods referred to in phase 2 with the results of the UT method. This additional analysis aimed to conduct a rigorous comparison and thus increase the quality of research. All the results in the previous section are reduced in this section as a result of conducting a more thorough comparison.

4.3.1. Number of Usability Problems and Severity Rating

All methods that were analyzed in Section 4.3 are compared with the UT method in this section in terms of the number of problems and their severity rating. The total number of unique problems discovered by all UEMs is 82 real problems. As presented in Table 4, the UES-HE method can identify 60.98% of the total number of real problems, and again it outperforms all UEMs, which is acceptable. The results of the UT and URS-HE methods are unsatisfactory because they discovered 52.44% and 43.90% of the total number of real problems. However, the HE method presented the worst result, which identified approximately 30% of real problems in total.

In regard to the severity rating for each UEM, Table 5 contains the number of cosmetic, minor, major, and catastrophic usability problems for each of these methods. The UES-HE method tended

Table 3. Usability methods and severity rating of usability problems

	UES-HE Method		URS-HE Method		HE Method	
1: Cosmetic	15	30.0%	11	30.6%	8	33.3%
2: Minor	13	26.0%	9	25.0%	8	33.3%
3: Major	17	34.0%	12	33.3%	5	20.8%
4: Catastrophic	5	10.0%	4	11.1%	3	12.5%
Total real problem	50	100.0%	36	100.0%	24	100.0%

Table 4. Usability methods and discovered usability problems

	UES-HE Method	URS-HE Method	HE Method	UT Method
No. of discovered problems	50	36	24	43
Percentage of discovered problems	60.98%	43.90%	29.27%	52.44%

Table 5. Usability methods and severity rating of usability problems

	UES-HE Method		URS-HE Method		HE Method		UT Method	
1: Cosmetic	15	30.0%	11	30.6%	8	33.3%	10	23.3%
2: Minor	13	26.0%	9	25.0%	8	33.3%	11	25.6%
3: Major	17	34.0%	12	33.3%	5	20.8%	16	37.2%
4: Catastrophic	5	10.0%	4	11.1%	3	12.5%	6	14.0%
Total real problem	50	100.0%	36	100.0%	24	100.0%	43	100.0%

to find more cosmetic, minor, and major problems compared to the other methods, whereas the UT method discovered a larger number of catastrophic problems. Among all UEMs, the HE method presented the lowest number of all problem types.

4.3.2. Overlapping and Unique Problems

One of the most interesting recommendations in the usability field is the use of more than two UEMs to improve the results, especially of the HE and UT methods (Fernandez et al., 2011). This recommendation is based on the argument that each method can find unique real problems that cannot be discovered by other UEMs and that each method has advantages and disadvantages and therefore that each method can be used complementary to the other methods. In this section, this argument is examined to determine whether it should be accepted or rejected.

Table 6 lists the unique real problems that were discovered by each of the UEMs. The UT method identified two cosmetic, six minor, six major, and four catastrophic problems. Thus,

Table 6. Severity rating for each UEM

	Methods	Cosmetic	Minor	Major	Catastrophic	Total Usability Issues	
Unique real problem	UES-HE	4	2	1	1	8	9.8%
	URS-HE	2	2	1	0	5	6.1%
	HE	3	1	1	0	5	6.1%
	UT	2	6	6	4	18	22.0%
Total overlapping real problem	UES-HE and HE	5	7	4	3	19	23.2%
	UES-HE and UT	5	5	9	2	21	25.6%
	URS-HE and HE	4	5	2	3	14	17.1%
	URS-HE and UT	4	3	4	2	13	15.9%
	UES-HE and URS-HE	6	7	10	4	27	32.9%
	HE and UT	0	2	2	2	6	7.3%
Total overlapping and unique real problem	UES-HE and HE	18	14	18	5	55	67.1%
	UES-HE and UT	20	19	24	9	72	87.8%
	URS-HE and HE	15	12	15	4	46	56.1%
	URS-HE and UT	17	17	24	8	66	80.5%
	UES-HE and URS-HE	20	15	19	5	59	72.0%
	HE and UT	18	17	19	7	61	74.4%

it found more minor, major, and catastrophic problems than the other methods. The UES-HE method found four cosmetic, two minor, one major, and one catastrophic problem; hence, it provides more satisfactory results in terms of cosmetic problems. The URS-HE method found two cosmetic, two minor, and one major problem, whereas the HE method found three cosmetic, one minor, and one major problem. However, neither of the latter two methods succeeded in finding a catastrophic problem.

Figure 3 illustrates the overlapping and unique real problems identified by all UEMs. The UT method found more unique problems than the other methods with 22.0% of the problems not found by the other methods (18 of the total number of real problems). The next most satisfactory result can be seen with the UES-HE method that could find 9.8% unique real problems (eight of the total number of real problems). Following this, the URS-HE and HE methods both found the same number of unique real problems, that is, 6.1% (five of the total number of real problems). Figure 3 also illustrates that the use of UES-HE with UT methods obtained superior results 87.8% (72 of 82). In addition to finding a greater number of problems overall, they were also superior in terms of the type of problems in that they identified 20 cosmetic, 19 minor, 24 major, and 9 catastrophic over and above those found by other methods. Moreover, using both the URS-HE and UT methods can release 80.5% (66 of a total of 82 real problems). Using the HE and UT methods also offer optimal results with 74.4% (61 out of a total of 82 real problems). The result of combining the methods are as follows: 72.0% for using the UES-HE and URS-HE methods, 67.1% for using the UES-HE and HE methods, whereas using URS-HE and HE offer the worst results at 56.1%. In summary, it can be argued that using UES-HE in combination with UT methods produces more optimal results. In addition, combining the UT method with each of the UES-HE, URS-HE, or HE methods can find an adequate number of usability problems. Therefore, these results support the recommendation that using more than one UEM offers improved results and confirm that no single method alone can identify the most usability problems.

4.3.3. UEMs Performance Metrics

4.3.3.1. Validity

Calculation of the validity metric did not require us to determine the realness of the usability problems. Table 7 presents the validity results. Both UES-HE and URS-HE methods provide satisfactory results at 80.6% and 83.7%, respectively. That means both of the proposed methods have acceptable validity and there are no significant false problems. However, the HE method was prone to false problems and offered unsatisfactory results at 58.5% (see Figure 4).

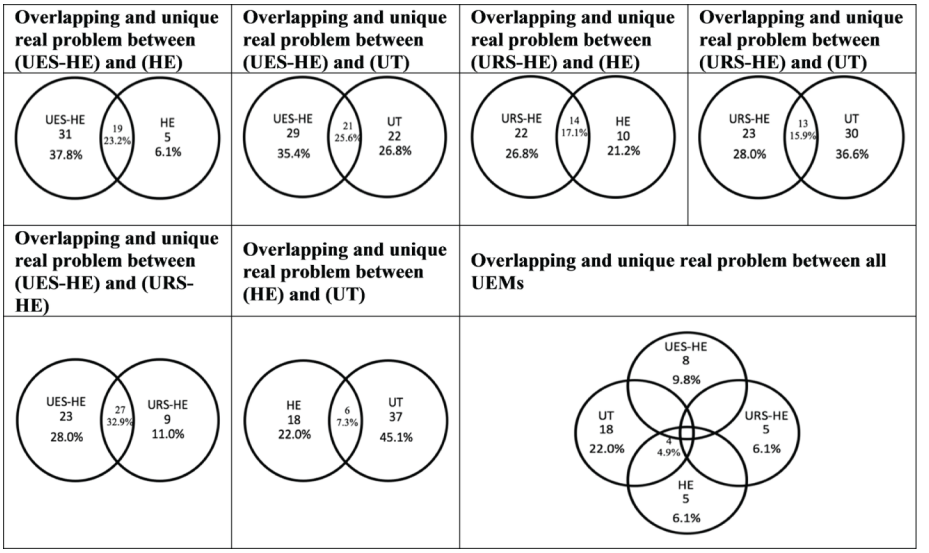
4.3.3.2. Thoroughness

Measuring the thoroughness of each method can be achieved by dividing the total number of real problems identified by a specific method by the real problems discovered by all the methods. That means the thoroughness required us to determine the realness of usability problems. Table 8 shows that

Table 7. Validity of usability evaluation methods

Methods	Total Usability Issues	0: Not a Problem		1: Cosmetic		2: Minor		3: Major		4: Catastrophic		Total Real Usability Problems	Validity	
UES-HE	62	12	19.4%	15	24.2%	13	21.0%	17	27.4%	62	12	19.4%	15	24.2%
URS-HE	43	7	16.3%	11	25.6%	9	20.9%	12	27.9%	43	7	16.3%	11	25.6%
HE	41	17	41.5%	8	19.5%	8	19.5%	5	12.2%	41	17	41.5%	8	19.5%
UT	43	0	0.0%	10	23.3%	11	25.6%	16	37.2%	43	0	0.0%	10	23.3%

Figure 4. Venn diagrams of the overlapping and unique real problems found by each UEM



the UES-HE method can offer satisfactory thoroughness whereas the result of the URS-HE method was unsatisfactory. The thoroughness of HE method was the worst at 29.3% as shown in Table 8.

4.3.3.3. Effectiveness

After obtaining the validity and thoroughness of all the UEMs, the effectiveness of each usability evaluation method can be conducted because the effectiveness is the output of multiplying the validity with the thoroughness.

Among all UEMs, as shown in Table 9, the UES-HE method outperformed all the other methods in terms of effectiveness. The results of the other methods are unsatisfactory, where the effectiveness result of the URS-HE method showed that this method identified half of the total number of usability

Table 8. Thoroughness of usability evaluation methods

Methods	Real usability Problems	Total Real Usability Problems	Thoroughness	
UES-HE	50	82	0.61	61.0%
URS-HE	36		0.44	43.9%
HE	24		0.29	29.3%
UT	43		0.52	52.4%

Table 9. Effectiveness of usability evaluation methods

Methods	Validity		Thoroughness		Effectiveness	
UES-HE	0.81	80.6%	0.61	61.0%	0.49	49.4%
URS-HE	0.84	83.7%	0.44	43.9%	0.37	37.0%
HE	0.59	58.5%	0.29	29.3%	0.17	17.1%
UT	1	100.0%	0.52	52.4%	0.52	52.0%

problems existing in the system (49.4%) and the effectiveness of the HE method was again the worst effectiveness result with 17.1%.

The results in this section clearly show that there is a variation in results obtained by the different UEMs. UES-HE outperformed the other methods in terms of the number of usability problems. The reason for this may be that this method benefited from the exploratory session that allows the evaluator to understand the end-user. However, the UES-HE method was more prone to false problems than the URS-HE method and that may be attributable to the fact that it does not permit evaluators to review and discuss their list of problems with end-users, unlike the URS-HE method. URS-HE outperformed the HE method and this superiority may be because the evaluator is allowed to benefit from reviewing and discussing their result with the end-user. One striking result is that the HE method presents the worst results among the UEMs in terms of the number of usability problems and usability performance metrics. This may be ascribed to its failure to involve the end-user, unlike the proposed methods. Furthermore, it shows more false problems than the other methods. This again confirms that the traditional HE method remains prone to false problems and should be examined by researchers to alleviate this tendency. Although the UT method offers the second most satisfactory results in terms of the number of usability problems after UES-HE, it was more effective than the other methods. The UT method may be distinguished on the basis of its dependence on the direct observation of the end-user, but it is limited to the use tasks.

5. CONCLUSION

Heuristic evaluation is a widely accepted inspection method for diagnosing potential usability problems, and it can be used by expert and non-expert evaluators alike. Although the former type of evaluator yields enhanced results compared to the latter, experts are hard to find. To improve the results obtained by non-expert evaluators, this research proposed the introduction of two user sessions within the HE process. The outcome of this paper showed that non-expert evaluators within the HE method should be encouraged to include end-users during their evaluation activity because this has been proven to be more effective than the traditional method. Involving end-users can play a vital role in the results obtained with both of the proposed methods (UES-HE and URS-HE). Both of the proposed methods yielded a more satisfactory result than the traditional HE method in terms of effectiveness and the number of usability problems identified. These results confirmed the influence of involving the end-user in the HE method and thus future research should focus on the role of the end-user in this method. Although the UES-HE method discovered more usability problems than other methods (URS-HE, HE, and UT methods), it is not more efficient than the UT method. However, the result shows that none of the UEMs alone can offer comprehensive coverage of all usability problems that were discovered in the system. In addition, this research is in line with research that suggests using two evaluation methods to obtain satisfactory results. Using (UES-HE) and (UT) in combination can discover 87.8% of the total number of real problems in the system.

This study shows that the traditional method of relying on non-experts does not give desired results. Therefore, we conclude from the foregoing that the proposal to involve the user with the non-expert evaluator can improves non-expert performance, compared to the traditional method that relies solely on the expert. These proposed methods can provide developers with an alternative method to the traditional method of HE. Also, the findings of this study can provide an alternative method for developers in case of difficulty finding an expert evaluator or in case that their recruitment is costly. In addition, the proposed methods contribute to improving the understanding of the field of usability, by researching the knowledge gap mentioned in the Introduction section. This is important because every day there is rapid growth in the number of websites in the world; and because of the difficulty of obtaining experts to evaluate this growing number of websites, there is a need to find alternative methods that do not

depend on experts, yet achieve satisfactory results. Therefore, such a study could shift the focus of future research towards alternative methods that can be described as, 'discount methods', because of the ease of availability of non-expert evaluators and users. However, the results of this research suggest the need for further investigation by combining the two proposed methods in this research into one method named UERS-HE. Thus, we plan to examine the possibility of using this method in future research. In addition, further experiments can be conducted in future by analyzing the UERS-HE method in terms of the different designs of collaborative heuristic evaluation (CHE).

REFERENCES

Äijö, R., & Mantere, J. (2001). Are non-expert usability evaluations valuable? *Paper presented at the 18th International Symposium on Human Factors in Telecommunications (HfT 2001)*, Bergen, Norway.

Allen, M., Currie, L. M., Bakken, S., Patel, V. L., & Cimino, J. J. (2006). Heuristic evaluation of paper-based Web pages: A simplified inspection usability methodology. *Journal of Biomedical Informatics*, *39*(4), 412–423. doi:10.1016/j.jbi.2005.10.004 PMID:16321575

Alqurni, J., & Pooley, R. (2016). The influence of end-users on non-expert evaluator output within the heuristic usability method. *International Review of Basic and Applied Sciences*, *4*(3).

AlRoobaea, R., Al-Badi, A., & Mayhew, P. (2013). Generating a domain specific inspection evaluation method through an adaptive framework: a comparative study on educational websites. *International Journal of Human-Computer Interaction*, *4*(2), 88.

Baker, K., Greenberg, S., & Gutwin, C. (2001). Heuristic Evaluation of Groupware Based on the Mechanics of Collaboration. In M. Little & L. Nigay (Eds.), *Engineering for Human-Computer Interaction* (Vol. 2254, pp. 123–139). Springer Berlin Heidelberg. doi:10.1007/3-540-45348-2_14

Botella, F., Alarcon, E., & Peñalver, A. (2013). A new proposal for improving heuristic evaluation reports performed by novice evaluators. *Paper presented at the Proceedings of the 2013 Chilean Conference on Human - Computer Interaction*, Temuco, Chile. doi:10.1145/2535597.2535601

Chen, C. (2012). A usability inspection expert system based on HE, GRY and GST. *International Journal of Intelligent Information Processing*, *3*(1), 1–15. doi:10.4156/ijiip.vol3.issue1.1

Chen, S. Y., & Macredie, R. D. (2005). The assessment of usability of electronic shopping: A heuristic evaluation. *International Journal of Information Management*, *25*(6), 516–532. doi:10.1016/j.ijinfomgt.2005.08.008

Cheng, L. C. & Mustafa, M. (2014). A Reference to Usability Inspection Methods.

Desurvire, H., & Thomas, J. C. (1993). Enhancing the performance of interface evaluators using non-empirical usability methods. *Paper presented at the Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. doi:10.1177/154193129303701702

Dumas, J. S., & Redish, J. (1999). *A practical guide to usability testing*. Intellect Books.

Edwards, P. J., Moloney, K. P., Jacko, J. A., & Sainfort, F. (2008). Evaluating usability of a commercial electronic health record: A case study. *International Journal of Human-Computer Studies*, *66*(10), 718–728. doi:10.1016/j.ijhcs.2008.06.002

Faulkner, L. (2003). Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments, & Computers*, *35*(3), 379–383. doi:10.3758/BF03195514 PMID:14587545

Fernandes, P., Conte, T., & Bonif'cio, B. (2012). WE-QT: A Web Usability Inspection Technique to Support Novice Inspectors. doi:10.1109/sbes.2012.30

Fernandez, A., Insfran, E., & Abrahão, S. (2011). Usability evaluation methods for the web: A systematic mapping study. *Information and Software Technology*, *53*(8), 789–817. doi:10.1016/j.infsof.2011.02.007

Fu, L., Salvendy, G., & Turley, L. (2002). Effectiveness of user testing and heuristic evaluation as a function of performance classification. *Behaviour & Information Technology*, *21*(2), 137–143. doi:10.1080/02699050110113688

Gray, W. D., & Salzman, M. C. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction*, *13*(3), 203–261. doi:10.1207/s15327051hci1303_2

Group, M. M. (2016). Internet World Stats. Retrieved from http://www.internetworldstats.com/

Hartson, H. R., Andre, T. S., & Williges, R. C. (2003). Criteria for evaluating usability evaluation methods. *International Journal of Human-Computer Interaction*, *15*(1), 145–181. doi:10.1207/S15327590IJHC1501_13

Hollingsed, T., & Novick, D. G. (2007). Usability inspection methods after 15 years of research and practice. *Paper presented at the 25th annual ACM international conference on Design of communication*, El Paso, TX. doi:10.1145/1297144.1297200

Holzinger, A. (2005). Usability engineering methods for software developers. *Communications of the ACM*, *48*(1), 71–74. doi:10.1145/1039539.1039541

Howarth, J., Smith-Jackson, T., & Hartson, R. (2009). Supporting novice usability practitioners with usability engineering tools. *International Journal of Human-Computer Studies*, *67*(6), 533–549. doi:10.1016/j.ijhcs.2009.02.003

Hwang, W., & Salvendy, G. (2007). What makes evaluators to find more usability problems?: A meta-analysis for individual detection rates. *Human-Computer Interaction, Pt 1. Proceedings*, *4550*, 499–507.

ISO. (1998). 9241-11 Ergonomic requirements for office work with visual display terminals (VDTs) - Part 11: Guidance on usability: ISO.

Kennedy, S. (1989). Using video in the BNR usability lab. *ACM SIGCHI Bulletin*, *21*(2), 92–95. doi:10.1145/70609.70624

Khajouei, R., Hasman, A., & Jaspers, M. W. M. (2011). Determination of the effectiveness of two methods for usability evaluation using a CPOE medication ordering system. *International Journal of Medical Informatics*, *80*(5), 341–350. doi:10.1016/j.ijmedinf.2011.02.005 PMID:21435943

Koutsabasis, P., Spyrou, T., Darzentas, J. S., & Darzentas, J. (2007). On the performance of novice evaluators in usability evaluations. In *Proc. PCI, Patra*.

Lindgaard, G., & Chattratichart, J. (2007). Usability testing: what have we overlooked? *Paper presented at the SIGCHI Conference on Human Factors in Computing Systems*, San Jose, CA. doi:10.1145/1240624.1240839

Muller, M. J., Matheson, L., Page, C., & Gallup, R. (1998). Methods & tools: participatory heuristic evaluation. *interactions, 5*(5), 13-18.

Nielsen. J. (2001). Did Poor Usability Kill E-Commerce? Retrieved from http://www.nngroup.com/articles/did-poor-usability-kill-e-commerce/

Nielsen, J. (1992). Finding usability problems through heuristic evaluation. *Paper presented at the SIGCHI Conference on Human Factors in Computing Systems*, Monterey, CA.

Nielsen, J. (1993). *Usability Engineering*. Morgan Kaufmann Publishers Inc.

Nielsen, J. (1994a). Enhancing the explanatory power of usability heuristics. *Paper presented at the SIGCHI Conference on Human Factors in Computing Systems*, Boston, M USA.

Nielsen, J. (1994b). Heuristic evaluation. In N. Jakob & L. M. Robert (Eds.), Usability inspection methods (pp. 25-62). John Wiley & Sons, Inc.

Nielsen, J. (1999). User interface directions for the Web. *Communications of the ACM*, *42*(1), 65–72. doi:10.1145/291469.291470

Nielsen, J. (2000). Why you only need to test with 5 users. Retrieved from https://www.nngroup.com

Nielsen, J. (2006). Quantitative Studies: How Many Users to Test? Retrieved from https://www.nngroup.com

Oracle. (2012). FAQ: How to conduct Heuristic Evaluation.

Osterbauer, C., Kφhle, M., Grechenig, T., & Tscheligi, M. (2000). Web Usability Testing: A case study of usability testing of chosen sites (banks, daily newspapers, insurances). In *the Sixth Australian World Wide Web Conference*.

Paz, F., Paz, F. A., Villanueva, D., & Pow-Sang, J. A. (2015, April 13-15). Heuristic Evaluation as a Complement to Usability Testing: A Case Study in Web Domain. *Paper presented at the 2015 12th International Conference on Information Technology - New Generations (ITNG)*.

Preece, J., Sharp, H., & Rogers, Y. (2015). Interaction design: beyond human-computer interaction (4th ed.). Chichester: John Wiley & Sons.

Rubin, J., & Chisnell, D. (2008). *Handbook of usability testing: How to plan, design, and conduct effective tests*. John Wiley & Sons.

Sears, A. (1997). Heuristic walkthroughs: finding the problems without the noise. *International Journal of Human-Computer Interaction*, *9*(3), 213–234. doi:10.1207/s15327590ijhc0903_2

Sherman, P. (2009). *UPA Salary Survey*. Illinois, US: Usability Professionals' Association.

Shneiderman, B. (1998). Designing the user interface: strategies for effective human-computer interaction (3rd ed.). Addison-Wesley.

Slavkovic, A., & Cross, K. (1999). Novice heuristic evaluations of a complex interface. *Paper presented at the CHI '99 Extended Abstracts on Human Factors in Computing Systems*, Pittsburgh, PA. doi:10.1145/632716.632902

Sova, D. H., & Nielsen, J. (2003). *234 Tips and Tricks for Recruiting Users as Participants in Usability Studies*.

Turner, C. W., Lewis, J. R., & Nielsen, J. (2006). Determining usability test sample size. *International encyclopedia of ergonomics and human factors, 3*, 3084-3088.

van den Haak, M. J., & de Jong, M. D. (2005). Analyzing the interaction between facilitator and participants in two variants of the think-aloud method. *Paper presented at the Professional Communication Conference IPCC 2005*. doi:10.1109/IPCC.2005.1494192

Van den Haak, M. J., de Jong, M. D. T., & Schellens, P. J. (2004). Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: A methodological comparison. *Interacting with Computers*, *16*(6), 1153–1170. doi:10.1016/j.intcom.2004.07.007

Virzi, R. A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *34*(4), 457–468. doi:10.1177/001872089203400407

Woolrych, A., Cockton, G., & Hindmarch, M. (2005). Knowledge resources in usability inspection. In *Proceedings of the HCI 2005*.

Zaharias, P., & Koutsabasis, P. (2012). Heuristic evaluation of e-learning courses: A comparative analysis of two e-learning heuristic sets. *Campus-Wide Information Systems*, *29*(1), 45–60. doi:10.1108/10650741211192046

Zhijun, Z. (2007). Usability Evaluation. In Z. Panayiotis & K. Sri (Eds.), *Human Computer Interaction Research in Web Design and Evaluation* (pp. 209–228). Hershey, PA: IGI Global.

Jehad Alqurni is currently pursuing the PhD degree in Computer Sciences with Heriot-Watt University, School of Mathematical and Computer Sciences, United Kingdom. His research interests include usability engineering, user experience, systems development and human–computer interaction.

Roobaea Alroobaea is currently Assistant Professor in College of Computers and Information Technology, Taif University, Kingdom of Saudi Arabia. He received a PhD degree in Computer Science in 2016, from University of East Anglia (UK) and the master's degree in Information System from University of East Anglia (UK) in 2012. He achieved a distinction in his bachelor's degree in computer science from King Abdulaziz University (KAU) in Kingdom of Saudi Arabia, in 2008. He is a Chair of support researches and system at Deanship of scientific research in Taif University. He has been honoured by HRH Prince Mohammed bin Nawaf Al Saud, the Saudi ambassador to the UK, in recognition of his research excellence at the University of East Anglia. His research interests include Human Computer Interaction, Cloud Computing and Machine Learning

Mohammed Alqahtani he was a Teaching Assistant with the Imam Abdulrahman Bin Faisal University, Dammam, Kingdom of Saudi Arabia from 2004 to 2009. In the same university he worked as a Cisco instructor from 2005 to 2009, as a lecturer from 2009 to 2015, and as an Assistant Professor since 2016. Currently he is the Chairman of the Computer Information System Department at the same university. His research interest includes the Software Engineering, Programming, Mobile Transactions, Usability and Data Analysis.